Morphology

Formele en Natuurlijke Talen

Lecture 6

Agenda

- What is the structure of words?
- How does this structure vary across languages?
- How can finite-state machines be used to analyze this structure?

These slides based on work by Jakub Dotlačil, Rick Nouwen, and Lori Levin

What is a word?

- What's in between spaces?
  - Dutch *kinderopvangtoeslagaffaire*
  - English *childcare benefits scandal*
  - Dutch *kinderopvangtoeslagaffaire*
  - English *childcare benefits scandal*
  - Vietnamese *cà phê* 'coffee'
  - Greenlandic *anartarfilerisuupput* 'they are the sewage collectors'
- What expresses a certain meaning?
  - koek / koeken / koekje / koekjes
  - dansen / dans / danst / danste / gedanst
- Patterns of relatedness are **productive**:
  - googlen: google, googlet, googlen, googlede, gegoogled
  - sms: smsje, smsjes

Translation

chair	stoel
chairs	stoelen
ball	bal
balls	ballen
chest	kist
chests	kisten
...	...

chair	stoel
ball	bal
chest	kist
plural **-s**	plural **-en**

What are the atoms of language?

- Storing each form: costly, inefficient
- Generalizations suggest **rules**
  - een koek / *een koeken / een koekje / *een koekjes
  - een boek / *een boeken / een boekje / *een boekjes
  - *ik dansen / ik dans / *ik danst / ik danste
  - *ik bakken / ik bak / *ik bakt / ik bakte
- For example: [een$_{\text{sg}}$ koek$_{\text{sg}}$], [ik$_{\text{1sg}}$ bak$_{\text{1sg}}$]

⋆ ‘Words’ not atomic, but something smaller: *morphemes*

‘Words’ have structure

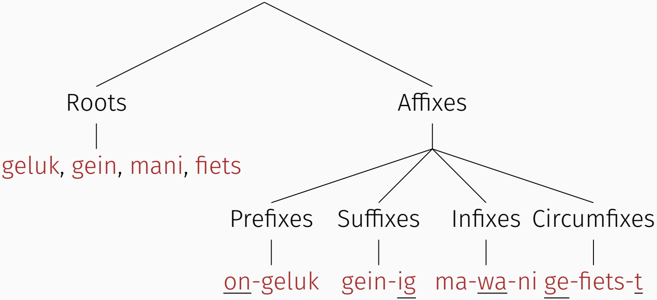
- **Morphemes**: smallest meaningful elements of a language
- Morphological *processes* combine morphemes to make larger units
  - boter
  - grot-er
  - chocolaa-tje-s
  - school-bord
  - ge-wandel-d
  - wandel-ing
- Bound morphemes: Can’t be used independently (e.g. *-en*)
- Free morphemes: Can be used independently (e.g. *kat*)

Morphological basics

Types of morphemes

Root: Carries the ‘core’ meaning of a word

Affixes: Serve derivational or inflectional functions (more on this later); attach to *stems* (root or root+affixes)



Roots

geluk, gein, mani, fiets

Affixes

Prefixes: on-geluk

Suffixes: gein-ig

Infixes: ma-wa-ni

Circumfixes: ge-fiets-t

Lakhota (Siouan; North/South Dakota):

mani ∼ walk

ma**wa**ni ∼ I walk

How morphemes are put together

Concatenation (prefixes, suffixes, circumfixes):

- stoel $\Rightarrow$ stoel-en
- steel $\Rightarrow$ steel-t
- stoel $\not\Rightarrow$ stoel-t

- Part of speech (noun, verb, etc.) matters!

Non-concatenative morphology

Apophony: Changing (but not adding) segments

- **Ablaut**: *foot* ~ *feet*
  sing ~ *sang* ~ *sung*
- **Initial consonant mutation** in Celtic languages

Welsh gender-based mutation (masculine vs. feminine)

Noun	Definite + Noun	Gloss
brawd	*y brawd*	‘(the) brother’
blodyn	*y blodyn*	‘(the) flower’
ffordd	*y ffordd*	‘(the) road’
merch	*y **f**erch*	‘(the) girl’
ryfel	*y **rh**yfel*	‘(the) war’
cwningen	*y **g**wningen*	‘(the) rabbit’

Non-concatenative morphology: suprasegmental

Suprasegmental morphology: Leaves segments (phonemes) the same but changes other aspects like stress or tone

Tone alternation Guébie (Kru; Ivory Coast) (Sande 2023)

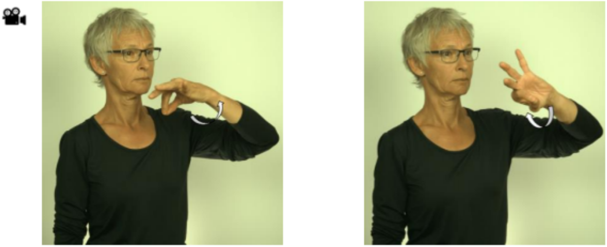
(1) a. ɟa^{31} nanɛ
coconuts be.good
‘coconuts are good’

b. ɟa$^{\mathbf{314}}$ nanɛ
coconuts.NEG be.good
‘coconuts are not good’

Suprasegmental morphology in signed languages

- Lots of grammaticized physical gestures: Eyebrow height, handshape, spatial position, direction, …
- Many options for suprasegmental inflection

Nederlands Gebarentaal (NGT) verbal agreement (Klomp 2021):



a. $_{1}$ANSWER$_{3a}$
‘I answer him/her/them.’

b. $_{3a}$ANSWER$_{1}$
‘She/he/they answer(s) me.’

Non-concatenative morphology: Reduplication

A root is fully or partially duplicated as an affix:

- **Hausa** (Chadic; W. Africa): intensification
  - can ‘there’
  - can-can ‘far away’
  - maza ‘fast’
  - maza-maza ‘very fast’
- **Yidiny** (Pama-Nyungan; Queensland, Aus): plurality
  - gindalba ‘lizard’
  - gindalgindalba ‘lizards’
- Compare: Ga je naar huis of naar huis-huis?
- Do you want a salad or a salad-salad?

Non-concatenative morphology: templates

Root-and-pattern/templatic morphology (Semitic langs: Arabic, Hebrew, Maltese, ...)

- 3-consonant roots (e.g. *k-t-b*)
- ‘Templates’ for word forms into which roots are inserted

Maltese Participle Stems:

Translation	Root	Perfect	Imperfect	Active
	C-C-C	CVCVC	VCCVC	CieCaC/CieCeC
‘to get cold’	k-s-ħ	kesaħ	eksaħ	kiesaħ
‘to sculpt’	n-q-x	naqax	onqox	
‘to ride’	r-k-b	rikeb	irkeb	riekeb

Types of morphological processes

The basics

- Three main processes for combining morphemes:
  - **Derivation**: Create new words from existing words: *un-happy, mogelijk-heid*
  - **Inflection**: Mark grammatically relevant features on words *walk-s, hog-e, boek-en*
  - **Compounding** (*samenstellen*): Combine two existing words: *fire-fighter, vis-handel*

Derivation

- Affixation of a bound morpheme
- Lexical class (noun, verb, etc.) usually changes
- Meaning (may) drastically change

wandel-ing, schrijv-er, computer-en, ver-grijz-en

blauw-ig, on-logisch, be-drinken

bemoeizucht-ig-heid

- Derivation is **not necessarily freely applicable**: *schrijv-ing, *on-verdrietig, *be-eten

Inflection

- Addition of bound morpheme (‘inflectional affix’)
- No change in lexical category
- No major change in meaning
- Often conditioned by specific syntactic environment
- Grammatically relevant (e.g. for agreement)
- Forms a **paradigm**

*deze boek / deze boek-en	inflection
*de man zijn …/ de mann-en zijn …	inflection
dit boek / dit boek-je	derivation
de gelukkige vrouw is …/ de on-gelukkige vrouw is …	derivation

Paradigm:

slaap	slap-en
slaap(-t)	slap-en
slaap-t	slap-en

Stem: part of the word to which morphemes are attached

Inflection on nouns

- Singular/Plural: boek / boek-en
- Case (*naamval*):

(2) German

Der	Mann	sieht	den	Sohn	des	Königs	in	dem	Garten
The$_1$	man	sees	the$_4$	son	the$_2$	king$_2$	in	the$_3$	garden

- Finnish partial case paradigm for talo 'house'

(3)

talo	talo-n	talo-na	talo-ksi	talo-ssa
nominative	accusative	partitive	translative	inessive

Inflection on adjectives

- Comparative form:
  - Dutch: slim => slimmer
  - English: smart => smarter
  - German: schlau => schlauer
- Not always across the whole lexicon:
  - More beautiful / *beautifuller
  - meer nodig / *nodiger
- Most languages: no special comparative morpheme

(4) Japanese

Nihongo-wa	doitsugo	yori	muzukashii
Japanese-TOP	German	from	difficult

'Japanese is more difficult than German.'

Verbal inflection

- Number: ik loop / wij lopen singular/plural
- Person: ik loop / hij loopt 1sg/3sg
- Tense: ik stap / ik stapte present/past

- Dutch vs. Slovenian:

1st	2nd	3rd	
maak	maakt	maakt	singular
maken	maken	maken	plural

1st	2nd	3rd	
delam	delaš	dela	singular
delava	**delata**	**delata**	**dual**
delamo	delate	delajo	plural

Inflection versus Derivation

Inflection doesn't change categories and takes places after derivation

- Tafel-tje-s / *Tafel-s-tje
- Wandel-ing / *Wandel-t-ing

Derivation, and not inflection, may be applied **recursively**:

industry
industri-al
industrial-ize
industrialize-ation
industrialization-al
industrializational-ize
...

Compounding

stem + stem

- Two content words pieced together to make new content word
- School-bord, tafel-kleed, achter-ingang, schaats-baan
- vries-drogen, zand-stralen
- sneeuw-wit, bloed-rood

- Can be distinguished prosodically from non-compounds:
  - hogeschool / hoge school
  - kleinkind / klein kind

Language typology

Language variation

- Languages differ dramatically in the kind and amount of morphology they make use of
- Dutch and English: relatively little inflection
- We can group languages based on their morphological tendencies, but these are not strictly delineated categories

Isolating (Analytic) languages

- Few (or no) bound morphemes (but extensive compounding)
- Larger role for word order and ‘function’ words

Mandarin:

wǒ = I	wǒ men = we
nǐ = you.SG	nǐ men = you.PL
tā = he/she	tā men = they.PL
rén = person	rén men = people

(5) Anhay da mua hai traicam. (Vietnamese)
he PAST buy two oranges
‘He bought two oranges.’

(6) khaw ca haj dek kin khaaw. (Thai)
he FUTURE CAUSE child eat rice

Agglutinative languages

- Extensive use of affixation
- Transparent meaning-morpheme relationships (1-1 feature-morpheme correspondedence)
- Examples: Hungarian, Finnish, Turkish

Turkish aorist (≈ future) paradigm:

gid-ér-im	gid-ér-sin	gid-ér-∅
go-AOR-1sg	go-AOR-2sg	go-AOR-3sg
gid-ér-iz	gid-ér-siniz	gid-ér-ler
go-AOR-1pl	go-AOR-2pl	go-3pl

Fusional languages

- Use single inflectional morphemes to reflect multiple grammatical categories (*portmanteau*)
- Examples: Slavic languages, Romance languages, German

1st	2nd	3rd	
povídá-m	povídá-š	povídá	singular
povídá-me	povídá-te	povída-jí	plural

Polysynthetic languages

Highly complex words, incorporating what might be a sentence in other languages, extensive inflection

(7) Inuktitut (Eskimo-Aleut; Canada) (Johns 2007)

annulaksi-kkanni-nginna-jualu-gasu-lauqsima-guma-nngit-tsiaq
imprison-again-really-a.lot-try-ever-want-NEG-EMPH-
-galuaq-tunga
EMPH-1SG.INTR.DECL
'I would never ever even want to try to end up in jail ever again even for a bit.'

Finite-state morphology

Is morphology regular?

Concatenative morphology: Mostly straightforwardly capturable with regular concatenation

Two wrinkles:

- Some processes have multiple (but predictable) realizations depending on properties of the root/stem
- Lexical exceptions (irregular words)

Example: English plural spelling

- *pizza-pizzas*, *oboe-oboes*, *wombat-wombats*
- *fox-foxes*, *bus-buses*, *city-cities*
- *goose-geese*, *child-children*, *mouse-mice*, *moose-moose*

Both of these caveats are easily addressed by finite-state means. (How?)

Morphological parsing

FSAs only **recognize**: telling us whether words are *legal*.

But we might also be interested in relation between word's form and its morphological structure:

- *wrote* = {write+V+PAST}

Relevant for **parsing** (analyzing a structure of a word given its form)

...and for **generation** (determining form given an analyzed structure)

We can do these by extending FSAs with rewrite power: finite-state *transducers*

Finite state transducers

A finite state transducer: $\langle \Sigma_1, \Sigma_2, S, s, A, R_1, R_2 \rangle$

Σ_1: input alphabet

Σ_2: output alphabet (**new!**)

S: states

$s \in S$: start state

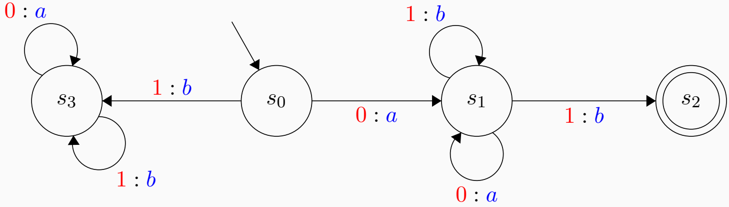
$A \subseteq S$: final states

$R_1 : (S \times \Sigma_1^*) \to S$: transition-relation

$R_2 : (S \times \Sigma_1^*) \to \Sigma_2^*$: output-relation (**new!**)

Input: A string from Σ_1

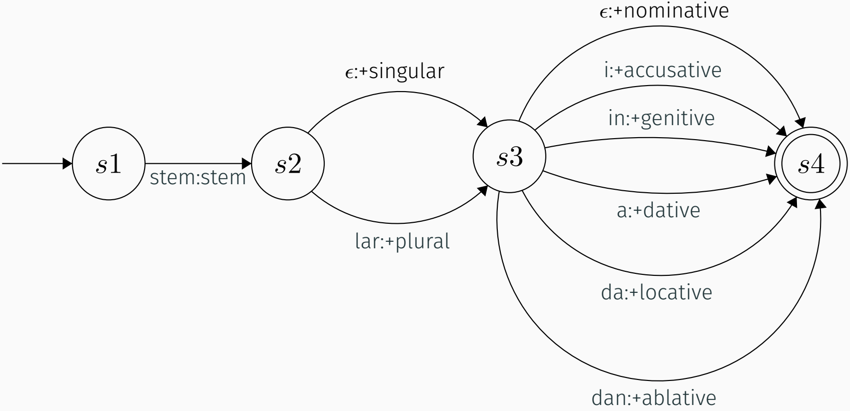
Output: A string from Σ_2



Application: parsing case in Turkish

Turkish	*singular*	*plural*
Nominative	adam	adamlar
Accusative	adami	adamlari
Genitive	adamin	adamlarin
Dative	adama	adamlara
Locative	adamda	adamlarda
Ablative	adamdan	adamlardan

An FST for parsing Turkish case



What kind of FST would we have if we swapped the inputs/outputs?

FSTs for translations

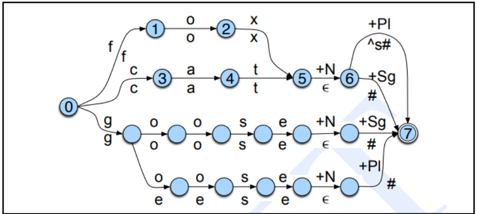


Figure 3.14 A fleshed-out English nominal inflection FST T_{lex}, expanded from T_{num} by replacing the three arcs with individual word stems (only a few sample word stems are shown).

FSTs can also be chained together, for purposes like cross-linguistic translation:

- English parser: chairs ⇒ chair+N+PL
- English-Dutch stem translator: chair+N+PL ⇒ stoel+N+PL
- Dutch generator: stoel+N+PL ⇒ stoelen

Computing non-concatenative morphology

Can non-concatenative morphology can be recognized by FSAs?

- Suprasegmentals: Yes, if we represent the suprasegmental component in the right way
- Infixation: Yes, by adding silent ”infix” marker into word representation
- Templatic morphology: Yes, but requires a lot of tricks

However...

One problem case

Bambara (Niger-Congo; Mali) reduplication

(8) wulu + **nyini** + **na** = wulunyinina
dog + search + for
'one who searches for dogs (dog searcher).'

(9) wulunyinina + **nyini** + **na** = wulunyininanyinina
dog searcher + search + for
'one who searches for dog searchers'

(10) wulunyinina + **O** + wulunyinina
dog searcher + of + dog searcher
'whichever dog searcher'

(11) wulunyininanyinina + **O** + wulunyininanyinina
dog searcher searcher + of + dog searcher searcher
'whoever searches for dog searchers'

Is Bambara reduplication regular?

The Pumping Lemma | For every regular language L there exists an integer p such that for every string $r \in L$ with $|r| \geq p$, there exist strings x, y and z such that:

- $r = xyz$
- $|xy| \leq p$
- $|y| > 0$
- For all $i \geq 0$: $xy^iz \in L$

$$A = \{\text{wulu(nyinina)}^n \text{ O wulu(nyinina)}^n | n \geq 0\}$$

- Assume that A is a regular language. Then there is a pumping length p such that all strings in A that are at least as long as p are pumpable.
- Let $r =$ wulu(nyinina)^p O wulu(nyinina)^p. ($|r| = 2p + 3 \geq p$)
- Two plausible ways to split r: $y =$ wulu or $y =$ (nyinina)^k where $0 < k \leq p - 1$.
- Now we pump: xy^iz, for example xy^2z
- xy^2z has too many *wulu* or too many *nyinina* before O, en therefore is not in A. r is not pumpable: Contradiction!

Are morphological processes regular?

- Bambara shows that some morphological processes result in non-regular languages

  ⇒ not *everything* in morphology is describable with finite-state means
- **But**: it's debatable whether these processes in Bambara are morphology *per se* or syntax (next week)
- Besides examples with unbounded copying, words built by morphological processes seem to be regular.

Summary

- Morphological processes: derivation, inflection, compounding
- Types of languages: analytic, agglutinative, fusional, polysynthetic
- FSAs and morphology: morphological processes by and large describable by FSAs
- Finite state transducers very useful for morphological parsing and generation